# Treatment Based Survival Prediction Using Data Mining Classification Algorithm on Breast Cancer

B. Lavanya [#1], Sri lakshmi R [#2]

[#] *Department of computer Science,*

*University of Madras*

[1] lavanmu@gmail.com

[2] srilakshmir074@gmail.com

***Abstract*— When a patient gets diagnosed with breast cancer, she is advised with specific treatment by the doctor. Every treatment is also associated with a probability of survival. Treatment will be suggested based on analysis of the cancer dataset. This dataset contains attributes such as age, place of residence (rural / urban), stage, tumor size, etc., Doctor will refer these data pertaining to the specific patient and predicts the survival probability for different treatment scenarios. Data mining is the method of finding or extracting information from massive databases or datasets and also contains several techniques and tools so that we can effectively collect and use the detection and prediction of medical data. This research compares the accuracy of data mining algorithms in predicting breast cancer survival.**

**Keywords – data mining, diagnosis, carcinoma, classification, prediction.**

## I INTRODUCTION

Healthcare industries are engendering and hoarding tremendous amount of data, which can be used to forecast and scrutinize the healthcare ratio of the entire country. It encompasses treatment, management and prevention of diseases based on patient's symptoms to address their health needs and preferences with the help of healthcare professionals using scientific medical knowledge including use of data mining (DM) techniques.

The rapid spread of carcinoma and therefore the inability to accurately diagnose and recognize its presence represents a challenge for researchers and developers in biomedical engineering. This challenge results in deploying new data processing techniques. Data mining is that the uprooting and recall of unknown data from the past which will be useful. Data mining also includes the acknowledged recovery and analysis of data that is saved in a data repository. Some of the important methods of knowledge mining are classification, association, clustering and regression, etc.

## II LITERATURE REVIEW

Literature survey describes the existing and established theory and research in your report area by providing a context for your work. This survey shows where you are filling a perceived gap in the existing theory or knowledge.

The automatic prediction of carcinoma is significant to abate the propensity against enlarging this disease. Early detection of carcinoma is the key to treatment. The study by Mosayebi et. al. [1] proposed a classification technique, namely stacking classifier for carcinoma prediction.

Alshammari et. al. [2] studied a common practical problem in the detection or recognition of data patterns using data mining techniques. In the WEKA tool with IBK algorithms, accuracy measurement is the highest. This paper analyses thirteen algorithms to find an accuracy value with elapsed time.

Loey et. al. research [3], based on IG feature selection, the GWO algorithm and SVM classification are used. Two microarray datasets were used as benchmarks to evaluate the 20 proposed methodology. The best results were obtained when combining the IG approach with both the GWO and SVM algorithms.

Keles et. al. [4] main aim was to predict cancer earlier and avoid biopsy using data mining techniques. The WEKA tool was used to detect carcinoma using data mining classification algorithms obtained from these attributes. All classification algorithms were tested and the most successful algorithms were determined based on accuracy rates. The high accuracy rates of these algorithms suggest that carcinoma can indeed be identified non-invasively, at low cost and without exposing patients to harmful radiation, by using data mining classification algorithms.

Salim Amour et. al presented in their paper with an approach for prediction and detection of carcinoma using machine learning techniques [5]. The presented tool can assist physician either new or experienced in diagnosis and prognosis of carcinoma at benign and malignant stages. Training and test results for carcinoma data sets are reported. The results show that the proposed DM disease prediction tool has potential to greatly impact on current patient management, care and future interventions against the carcinoma disease and through customization even against other deadly diseases.

This research work by Laghmati et. al. [6] presents an overview of some Machine learning techniques; ANN, KNN, Binary SVM, and decision tree. The four CAD techniques applied to the mammographic mass dataset; groups together 5 statistical attributes as input and a single binary output. The data analysis and classification supported the confusion matrix to extract the number of FN, TP, TN and FP. The results show that ANN is more reliable to assist in decision making regarding breast cancer severity.

This paper by khalshid et. al. [7] helps to improve the accuracy of breast cancer classification using data mining techniques. The UCI carcinoma dataset used and five data mining algorithms were used for the classification (Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), weighted K-Nearest Neighbors (Weighted KNN), and Gaussian Naïve Bayes (Gaussian NB) algorithms). The evaluation of results is done in terms of the confusion matrix and ROC curve.

The classification accuracy of detecting carcinoma is studied by Mohammed et. al. [8]. The three classifiers' algorithms J48, NB, and SMO were applied on two different carcinoma datasets.

Homogenous ensembles of Bagging and Boosting does not bring [9] about an increase in accuracy of algorithms in the

diagnosis of breast cancer while it increases the time taken by the algorithms to build its classification model. Bagging was shown to result in a slight increase for C4.5 and K-NN while it slightly reduced that of SVM while Boosting reduced classification accuracy in all cases.

[10] Zeinab Sajjadnia, Raof Khayami and Mohammad Reza Moosavi. Pre-processing Breast Cancer Data to Improve the Data Quality, Diagnosis Procedure, and Medical Care Services. Sage Journal. Cancer Informatics Volume 19: 1–16, Published: 2020.

## III DATASET:

This data set is obtained from the Hospital Based Cancer Registry (HBCR) Chennai. This includes 1709 new cases of female breast cancer that were registered and treated between 2010 and 2012, and are being followed up on until December 31, 2019.

### *A. Dataset Description:*

This data set has 30 distinct features. The term "reference number" (primary key) is used to identify the patient record uniquely serves as the primary key. Age, education, urban/rural area, marital status, religion, and mother tongue are the six variables that provide general information about the patient. Information about the test they took includes 11 factors. These include date of diagnosis, tumour type (histology), tumour grade, stage, nodal status, distant metastases, estragon receptor, progesterone receptor, and her2neu. Neo-adjuvant, adjuvant, surgery, radiation, chemotherapy, and hormonotherapy are the five aspects that provide treatment-related information. There are seven features that provide information regarding a patient's review. Second Cancer, Second Cancer Diagnosis Date, Disease-Free Survival, Disease-Free Survival Date, Recurrences, Overall Survival, and Overall Survival Date.

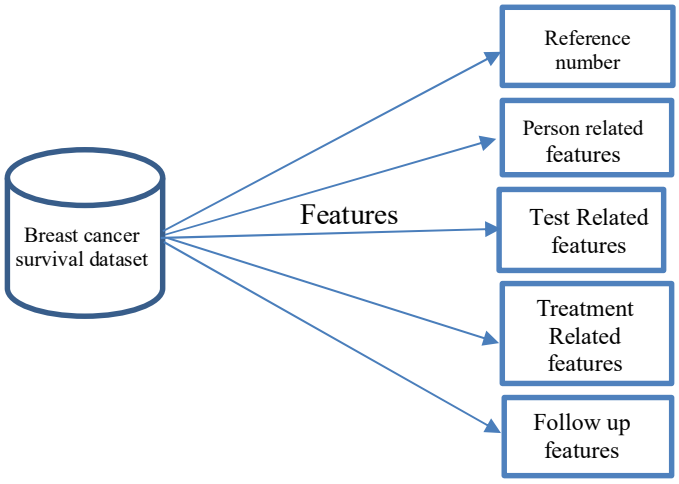


Fig 1: Data set structure

### *B. Feature selection for classification:*

The date set contains 30 attributes. such as domain expert advice factors that influence treatment-based survival categorization. From the data collection, 16 features are retrieved in fig 2.

| REFNO | age | stage 1 | t 1 | n 1 | m | hist 1 | grade 1 | er2 1 | pr2 1 | her22 1 | treatment code | sx | rt | ct | ht |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10092003041 | 34 | 2 | 2 | 1 | 0 | 1 | 3 | 0 | 0 | 1 | 2 | 1 | 1 | 1 | 0 |
| 10092003272 | 50 | 3 | 2 | 2 | 0 | 1 | 3 | 0 | 0 | 1 | 4 | 1 | 1 | 1 | 0 |
| 10092003338 | 39 | 2 | 2 | 1 | 0 | 1 | 3 | 0 | 0 | 9 | 2 | 1 | 1 | 1 | 0 |
| 10092003071 | 43 | 3 | 3 | 1 | 0 | 1 | 3 | 0 | 0 | 9 | 3 | 1 | 1 | 1 | 0 |
| 10092003072 | 29 | 3 | 3 | 2 | 0 | 1 | 3 | 0 | 0 | 0 | 5 | 0 | 1 | 1 | 0 |
| 10092003217 | 52 | 2 | 2 | 1 | 0 | 1 | 2 | 9 | 9 | 9 | 5 | 1 | 0 | 0 | 0 |
| 10092003029 | 48 | 2 | 2 | 1 | 0 | 1 | 3 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| 10092003282 | 50 | 3 | 3 | 2 | 0 | 2 | 9 | 9 | 9 | 9 | 4 | 1 | 1 | 1 | 0 |
| 10092004172 | 43 | 3 | 4 | 2 | 0 | 1 | 3 | 0 | 0 | 1 | 4 | 1 | 1 | 1 | 0 |
| 10092002419 | 33 | 2 | 2 | 1 | 0 | 2 | 3 | 0 | 0 | 0 | 2 | 1 | 1 | 1 | 0 |
| 10092003023 | 30 | 2 | 2 | 1 | 0 | 1 | 3 | 0 | 0 | 0 | 3 | 1 | 0 | 1 | 0 |
| 10092003090 | 55 | 3 | 4 | 1 | 0 | 1 | 3 | 0 | 0 | 1 | 4 | 1 | 1 | 1 | 0 |
| 10092003829 | 54 | 1 | 1 | 1 | 0 | 1 | 3 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| 10092002865 | 45 | 4 | 3 | 1 | 1 | 1 | 3 | 0 | 0 | 1 | 4 | 1 | 1 | 1 | 0 |
| 10092002820 | 43 | 1 | 1 | 1 | 0 | 1 | 3 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| 10092002980 | 48 | 3 | 3 | 2 | 0 | 1 | 3 | 0 | 0 | 1 | 4 | 1 | 1 | 1 | 0 |
| 10092003140 | 40 | 2 | 2 | 1 | 0 | 1 | 3 | 0 | 0 | 1 | 4 | 1 | 1 | 1 | 0 |
| 10092002963 | 55 | 3 | 2 | 2 | 0 | 1 | 3 | 0 | 0 | 1 | 4 | 1 | 1 | 1 | 0 |
| 10092003125 | 62 | 2 | 1 | 2 | 0 | 2 | 9 | 0 | 0 | 1 | 4 | 1 | 1 | 1 | 0 |
| 10092003539 | 56 | 2 | 2 | 1 | 0 | 1 | 2 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 10092003284 | 61 | 3 | 3 | 1 | 0 | 1 | 3 | 9 | 9 | 9 | 2 | 1 | 1 | 1 | 0 |
| 10092002951 | 48 | 3 | 4 | 2 | 0 | 1 | 3 | 0 | 0 | 0 | 4 | 1 | 1 | 1 | 0 |
| 10092003575 | 40 | 2 | 2 | 1 | 0 | 2 | 3 | 0 | 0 | 1 | 3 | 1 | 0 | 1 | 0 |
| 10092003305 | 50 | 3 | 4 | 2 | 0 | 1 | 3 | 0 | 0 | 0 | 4 | 1 | 1 | 1 | 0 |
| 10092002933 | 37 | 3 | 4 | 1 | 0 | 1 | 3 | 0 | 0 | 0 | 4 | 1 | 1 | 1 | 0 |

Fig 2: Screenshot for dataset extracting features

### *C. Proposed Work:*

The goal of this study is to forecast how long a cancer patient will live after treatment. Excel was used to calculate the overall survival date (in months) from the dataset containing both diagnosis and overall survival dates.

| REFNO | dxdt | osdt | OS_Duration |
|---|---|---|---|
| 10092003041 | 20-Feb-11 | 05-Feb-19 | 97 |
| 10092003272 | 12-Aug-11 | 06-May-19 | 94 |
| 10092003338 | 04-Aug-11 | 24-Sep-18 | 87 |
| 10092003071 | 03-Nov-11 | 17-Jul-19 | 94 |
| 10092003072 | 10-Oct-11 | 27-Jun-14 | 33 |
| 10092003217 | 09-Feb-11 | 08-May-11 | 3 |
| 10092003029 | 26-Jan-11 | 12-Jun-13 | 29 |
| 10092003282 | 26-Jul-11 | 08-May-19 | 95 |
| 10092004172 | 27-Jun-12 | 25-Apr-18 | 71 |
| 10092002419 | 23-Mar-10 | 09-Jan-19 | 107 |
| 10092003023 | 14-Mar-11 | 30-Apr-13 | 26 |
| 10092003090 | 05-Nov-11 | 01-Apr-13 | 17 |
| 10092003829 | 24-Mar-12 | 22-Feb-19 | 84 |
| 10092002865 | 09-Aug-10 | 12-Mar-19 | 105 |
| 10092002820 | 04-Nov-10 | 12-Jul-19 | 106 |
| 10092002980 | 18-May-11 | 15-Aug-12 | 15 |
| 10092003140 | 25-Oct-11 | 17-Jul-19 | 94 |
| 10092002963 | 09-Apr-11 | 17-Jul-19 | 101 |
| 10092003125 | 17-Sep-11 | 17-Jul-19 | 95 |
| 10092003539 | 29-Oct-11 | 15-Jul-16 | 57 |
| 10092003284 | 23-May-11 | 19-Nov-18 | 91 |
| 10092002951 | 22-Dec-11 | 15-Jan-13 | 13 |
| 10092003575 | 19-Dec-11 | 19-Feb-19 | 87 |
| 10092003305 | 24-Sep-11 | 19-Jul-19 | 95 |
| 10092002933 | 20-Mar-10 | 11-Feb-11 | 11 |

Fig 3: Screenshot os_duration from dxdt and osdt

Using SPSS software, found the overall survival duration to vital status (os_1) in fig 4. The surviving period is divided into four categories:

| **Level** | **Duration** |
|---|---|
| Level 0 | alive or dead after 5 years (more than 60 months) |
| Level 1 | within 1 year (12 months) |
| Level 2 | 1 to 3 years (13 – 36 months) |
| Level 3 | 3 to 5 years (37 – 59 months) |

Table 1: Survival duration level

| OS_Duration | os_1 |
|---|---|
| 97 | 3 |
| 94 | 3 |
| 87 | 3 |
| 94 | 3 |
| 33 | 1 |
| 3 | 0 |
| 29 | 1 |
| 95 | 3 |
| 71 | 3 |
| 107 | 3 |
| 26 | 1 |
| 17 | 1 |
| 84 | 3 |
| 105 | 3 |
| 106 | 3 |
| 15 | 1 |
| 94 | 3 |
| 101 | 3 |
| 95 | 3 |
| 57 | 2 |
| 91 | 3 |
| 13 | 1 |
| 87 | 3 |
| 95 | 3 |
| 11 | 0 |
| 100 | 3 |

Fig 4: Screenshot for converting numeric variable in to level

## IV IMPLEMENTATION:

Data mining refers to the process of automatically obtaining data from large amounts of data. The term "data mining" (also known as "knowledge discovery") refers to the process of analysing data for various purposes and extracting relevant information using a variety of tools and techniques, with the goal of improving a system's performance. It has developed prediction models by uncovering hidden patterns and studying the relationships between various sorts of data.

This study uses six data mining classification algorithms to predict classification of breast cancer dataset - Decision tree, Random Forest, SVM, GaussianNB and logistic regression. Classification accuracy is generally calculated by the percentage of instances correctly classified. The outcome of the data values given in the data set is included in the output. The accuracy of each categorization method was compared in the results.

### *A. Decision Tree:*

The supervised learning category includes the decision tree method. It can solve both regression and classification problems. It solves the problem using the tree representation, in which each leaf node corresponds to a class label and characteristics are represented on the tree's interior node. It divides a large data collection into smaller parts while also continuously developing the DT. At least two branches exist in a decision node. A classification can be seen in the leaf nodes. The root node is the highest decision node in a tree that connects to the best predictor.

In Decision Tree, the major challenge is in identification of the attribute for the root node in each level. This process is known as attribute selection. The two popular attribute selection measures: Information Gain and Gini index

Information Gain:

It refers to the decline in entopy after the dataset is split. It is also called entropy reduction. It partitions the training instances into smaller subsets of the entropy changes.

$$\text{Gain (S,A)} = \text{Entropy (S)} - \sum \text{values (A)} \frac{|Sv|}{|S|}$$

Entropy:

It refers to a common way to measure impurity. In DT, it measures the randomness or impurity in datasets.

It can be defined as a measure of purity of the sub split. Entropy always lies between 0 to 1.

$$H(s) = \sum_{i=1}^{c} -P_i \, log_2 \, Pi$$

Gini Index:

It is also working like entropy in DT. Both entropy and Gini are used for building the tree by splitting as per the appropriate features. But it has quite a difference in the computation part.

$$GI = 1 - \sum_{i=1}^{n} (p)^2$$

### *B. Random Forest Classifier:*

The Random Forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a randomly selected subset of the training set and then it collects the votes from different decision trees to decide the final prediction.
Random Forest Prediction for a classification problem: f(x) = majority vote of all predicted classes over B trees.

### *C. Support Vector Machine (SVM):*

SVM algorithm is a simple yet powerful Supervised machine learning algorithm that can be used for building both regression and classification models. It can be performed really well with both linearly separable and non-linearly separable datasets. This algorithm is based on the concept of 'decision planes', where hyper planes are used to classify a set of given objects.

SVM libraries are packed with some popular kernels such as polynomial, radial basis, function or rbf and sigmoid. The classification function used in SVM in machine learning is SVC. The SVC function looks like

SKlearn. SVM. SVC (C=1.0, Kernal='rbf', degree=3)

C – Keeping large values of C will indicate the SVM model to choose a smaller margin hyperplane.

Kernel – It is kernel type to be used in SVM model building. It can be 'linear', 'rbf', 'poly' or 'sigmoid'. The default value of kernel is 'rbf'.

Degree – It's only considered in the case of polynomial kernel. It is the degree of the polynomial kernel function. The default value of degree is 3.

### *D. Naive Bayesian Classification:*

Bayesian classifier is a statistical classifier. They can predict class membership probabilities, for instance, the probability that a given sample belongs to a particular class. Bayesian classification is created on the Bayes theorem. Studies comparing the

classification algorithms have found a simple Bayesian classifier known as the naive Bayesian classifier to be comparable in performance with decision tree and neural network classifiers. Bayesian classifiers have also displayed high accuracy and speed when applied to large databases. Naive Bayesian classifiers adopt that the exact attribute value on a given class is independent of the values of the other attributes. This assumption is termed class conditional independence. It is made to simplify the calculations involved, and is considered "naive". Bayesian belief networks are graphical replicas, which unlike naive Bayesian classifiers allow the depiction of dependencies among subsets of attributes. Bayesian belief can also be utilized for classification.

Bayes' Theorem:

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)}$$

*E. Logistic Regression:*

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable (or output), y, can take only discrete values for a given set of features (or inputs) X.

h(x) represents the predicted response for ith observation xi. The formula for calculating h(xi) is called hypothesis.

## V DATA ANALYSIS

The objectives of the data collection were to find out the prediction of overall survival for 5 years using treatment-based classification for breast cancer dataset. Using the classification models which prediction is gives best fit for this dataset.

*A. Data collection and its procedure:*

The Hospital Based Cancer Registry (HBCR) in Chennai provided this data collection. This comprises 1709 new cases of female breast cancer that were diagnosed and treated between 2010 and 2012 and are being tracked until December 31, 2019.

Finding the features that are suitable for this classification and extracting the features from the data set are the aims of this research. Because some of the features are not numerical, they must be converted to numerical values from coded values using SPSS software in order to be validated (cross validation) appropriately. Overall survival (5 years) is divided into four levels (levels 0–3) to provide the more beneficial for this research.

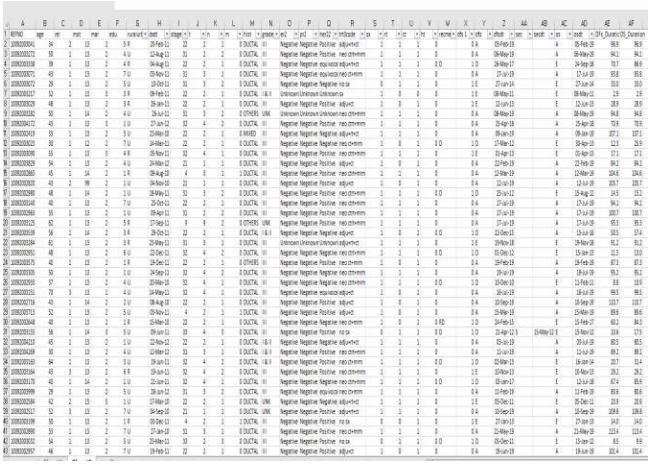

Fig: 5 Screen shot of the dataset which is collected

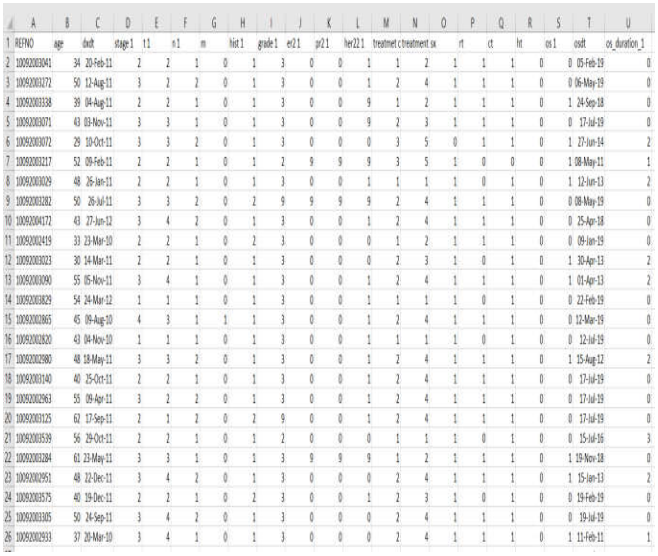

Fig: 6 Screen shot of the dataset for this classification

*B. Variable understanding:*

This data collection has 30 features, from which 16 were used for this study in fig6. The variable in the dataset which is more important to this analysis are

Treatment code:

Using this treatment code, to analyse which treatment gives best survival table:2.

| Treatment name | Coded for analysis |
|---|---|
| Adjuvant +chemotherapy | 1 |
| Adjuvant + chemotherapy; Adjuvant + chemotherapy + radiotherapy | 2 |
| Neo adjuvant chemotherapy + modified radical | 3 |
| Neo ct+ rt+ mrm; Neo rt + mrm | 4 |
| Other sx , no sx | 5 |

Table: 2 Treatment code recode for analysis

Using this, the study classifies the survival of the patient. In this study using 5 classification models to implementing it into the dataset.

Classification accuracy is generally calculated by the percentage of instances has correctly classified. The output includes the result of the data values given in the data set. The result was compared based on the Accuracy of each classification algorithms. This study has been performed using python with several data mining classification algorithms.

The breast cancer prediction system estimates risk of the breast cancers. This system was validated by comparing its predicted results with patient's prior medical information and analysed using python.

| Classification algorithm | No. of instance is used | Accuracy |
|---|---|---|
| Decision tree | 855 | 60% |
| Random forest | 428 | 68% |
| Naive bayes | 428 | 70% |
| SVM | 342 | 70% |
| Logistic regression | 513 | 72% |

Table:3 Accuracy level of different classifier algorithm

The Table 3 shows that the high accuracy belongs to different classification Algorithms. Based on this accuracy many algorithms show a better accuracy for a prediction.

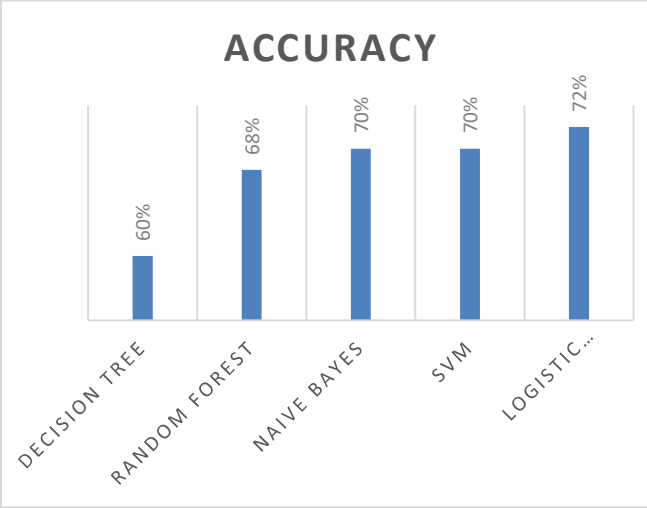


Fig 9: Accuracy of four different algorithms

The Fig 9 shows in overall, all algorithms have a better performance. But Logistic regression performance is little high when compared to other algorithms.

| Classification algorithm | Correctly classified | Incorrectly classified | accuracy |
|---|---|---|---|
| Decision tree | 513 | 342 | 60% |
| Random forest | 256 | 172 | 68% |
| Naive bayes | 300 | 128 | 70% |
| SVM | 243 | 99 | 70% |
| Logistic regression | 370 | 143 | 72% |

Table:4 Analysis of Different Classification algorithms

The Table 4 shows the different classifiers such as decision tree it has correctly classified 513 data, random forest it has correctly classified 256 data, NB algorithms it has correctly classified 300 data, SVM has 243 data, and Logistic regression has 370 data has correctly classified when compare to other algorithms, in these data set the Logistic regression shows better performance for this dataset.

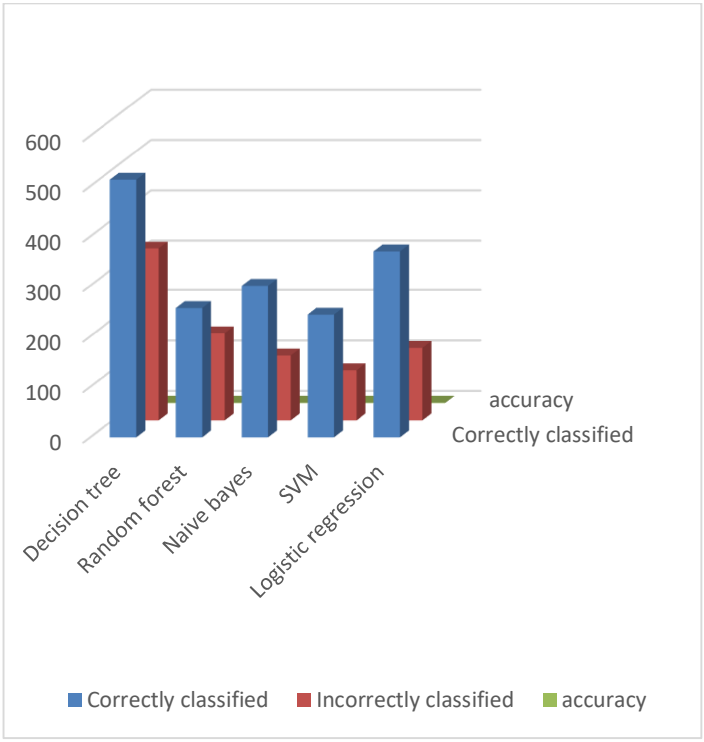


Fig 10: Performance of Different Classifier

The Figure 10 shows in overall all algorithms has a better performance. But logistic regression performance is little high when compared to other algorithms.

This analysis shows the data mining classification algorithms or more are less equally classified. From this dataset logistic regression gives higher accuracy than other classification models.

## VI CONCLUSION

Breast cancer is one of the major causes of death in women. In breast cancer, research ultimately improves the quality of Healthcare and Cancer patients. Using the treatment-based analysis which treatment for which stage of cancer patients give more survival chance of the patient. This study of classification techniques like Decision tree, Random Forest, Naïve bayes, SVM, logistic regression gives more or less same accuracy. From that the logistic regression gives more accuracy for this dataset. The future enhancement of this research work is to improve the accuracy implementing the artificial neural network classification in the dataset and find more accuracy.

## REFERENCE

[1] Alireza Mosayebi, Barat MojaradiI, Ali Bonyadi NaeiniI, Seyed Hamid Khodadad Hosseini. Modelling and comparing data mining algorithms for prediction of recurrence of breast cancer: Bryan C. Daniels, Arizona State University & Santa Fe Institute, UNITED STATES Published: October 15, 2020.

[2] Majdah Alshammari, Mohammad Mezher Department of Computer Science Fahad Bin Sultan University, Tabuk, KSA. A comparative analysis of data mining techniques on breast cancer diagnosis data using WEKA toolbox. (IJACSA) International Journal of Advanced Computer Science and Applications, page no. 224-229, Vol. 11, No. 8, 2020.

[3] Mohamed Loey Ramadan AbdElNabi, Mohammed Wajeeh Jasim, Hazem M. EL-Bakry, Mohamed Hamed N. Taha and Nour Eldeen M. Khalifa. Breast and Colon Cancer Classification from Gene Expression Profiles Using Data Mining Techniques. Symmetry 12, 408 Published: 4 March 2020.

[4] Mümine KAYA KELEŞ. Breast cancer prediction and detection using data mining classification algorithms: Technical Gazette 26, 1(2019), 149- 155. ISSN 1330-3651 (Print), ISSN 1848-6339 (Online).

[5] Salim Amour Diwani and Zaipuna Obedi Yonah. Holistic Diagnosis Tool for Early Detection of Breast Cancer. International Journal of Computing and Digital Systems. ISSN (2210-142X) Int. J. Com. Dig. Sys. 10, No.2 Apr-2021.

[6] Sara Laghmati, Amal Tmiri, Bouchaib Cherradi. Machine Learning based System for Prediction of Breast Cancer Severity. IEEE. 978-1-7281- 2625-8/19, Published: 2019.

[7] Shler Farhad Khorshid, Adnan Mohsin Abdulazeez and Amira Bibo Sallow. A Comparative Analysis and Predicting for Breast Cancer Detection Based on Data Mining Models. Asian Journal of Research in Computer Science. 8(4): 45-59, 2021; Article no. AJRCOS.68450 ISSN: 2581-8260 Published: 19 May 2021.

[8] Siham A. Mohammed, Sadeq Darrab, Salah A. Noaman, and Gunter Saake. Analysis of Breast Cancer Detection Using Different Machine Learning Techniques. Springer Nature Singapore Pte Ltd. Y. Tan et al. (Eds.): DMBD 2020, CCIS 1234, pp. 108–117, 2020.

[9] Taye Oladele Aro, Hakeem Babalola Akande, Muhammed Besiru Jibrin, Usman Abubakar Jauro. Homogenous ensembles on data mining techniques for breast cancer 41 diagnosis. Daffodil International University Journal of Science and Technology. Published: 1st July 2019.

[10] Zeinab Sajjadnia, Raof Khayami and Mohammad Reza Moosavi. Pre-processing Breast Cancer Data to Improve the Data Quality, Diagnosis Procedure, and Medical Care Services. Sage Journal. Cancer Informatics Volume 19: 1–16, Published: 2020.